2015 IEEE Wireless Communications and Networking Conference (WCNC) - Workshop - Energy Efficiency in the Internet of Things, and Internet of Things for Energy Efficiency

A non-intrusive model to predict the flexible energy in a residential building

Luc Dufour, Dominique Genoud, Antonio Jara, Jerôme Treboux[1]
Bruno Ladevie and Jean-Jacques Bezian [2]

1 Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
2 Mines Telecom, Albi, France
Email:(luc.dufour, dominique.genoud, antonio.jara, jerome.treboux)@hevs.ch,
(bruno.ladevie and bezian)@mines-albi.fr

***Abstract*—The building energy consumption represent 60% of total primary energy consumption in the world. In order to control the demand response schemes for residential users, it is crucial to be able to predict the different components of the total power consumption of a household. This work provide a non intrusive identification model of devices with a sample frequency of one hertz. The identification results are the inputs of a model to predict the flexible energy. This corresponds at the different devices could be shift in a predetermined time. In a residential building, the heating and the hot water represent this flexible energy. The Support Vector Machine (SVM) enable an identification around 95% of heating, hot water, household electrical and a ensemble of decision tree provide the prediction for the next 15 minutes.**

Keywords—Data intelligence analysis; Microgrid; Advanced Metering Infrastructure; Energy information management;KNIME;

I. Introduction

Non intrusive load monitoring (NILM) [4] is a general term which refers to determining the energy consumption of individual devices, or statistics of the energy consumption signal, without installing individual sensors at the plug level. The motivations for such a process are twofold. First, informing a households occupants of how much energy each appliance consumes empowers them to take steps towards reducing their energy consumption. Second, if the NILM system is able to determine the current time of use of each appliance, a system would be able to inform a households occupants and the distributor of the potential energy savings through deferring appliance use to a time of day when electricity is either cheaper or has a lower carbon footprint.
Moreover, they adapt over time in changes in households (such as changes in appliance number and type) without requiring new installations or reconfiguration of existing hardware and software. The extensive deployment of smart meters which is planned in many countries for the near future will enable a large scale deployment of NILM techniques [5]. Such deployment will make available measurements of the total active and reactive power consumed, typically sampled at low frequencies, allowing non-intrusive load monitoring without the use of additional hardware.
NILM methods have been first proposed in [6], and they are typically structured in three phases: feature extraction, events detection, and events classification. They make use of a database of electric signatures of appliances, and they are based on the measure of the total active power consumed, sampled at frequency of one Hertz. Later methods [7], [8] try to decrease the duration of the training period. Indeed, a fine granularity and a good accuracy in load disaggregation are crucial in order to enable useful feedback to users, to set up appropriate measures for changing consumption patterns, and to enable detection of anomalies and appliance malfunctioning. Many of the techniques proposed in order to overcome these drawbacks imply a substantially higher sampling frequency, and therefore more expensive hardware [9], [10].
This paper use a part of NILM method to identify the electrical devices, in particular the devices states. A Support vector machine enable the identification in the Active/Reactive power plan. A probabilistic neural network, a decision tree and a support vector machine provide the prediction.
The paper is organized as follows. In Section II the methodology with the identification and prediction model are presented. The Section III describe the setting used for the test and in Section IV we present the results of the different algorithms. Finally we conclude and discuss future directions of research in Section V.

II. Methodology

In this section, the identification and the prediction models are presented. The identification model used the active and reactive power events which are the target variables compute by a Cross-validation method [15], [16]. The predictive model used the output of the identification method to provide a level of flexibility for the next 15 minutes.

A. Training total Load Models

The total load depends on which appliances are switched on at any given moment, so we must describe a switch process, a(t). Suppose there are n appliances, numbered 1 to n and let a(t) be an n-component Boolean vector describing the state of the n switches at time t: 1,

978-1-4799-8760-3/15/$31.00 ©2015 IEEE

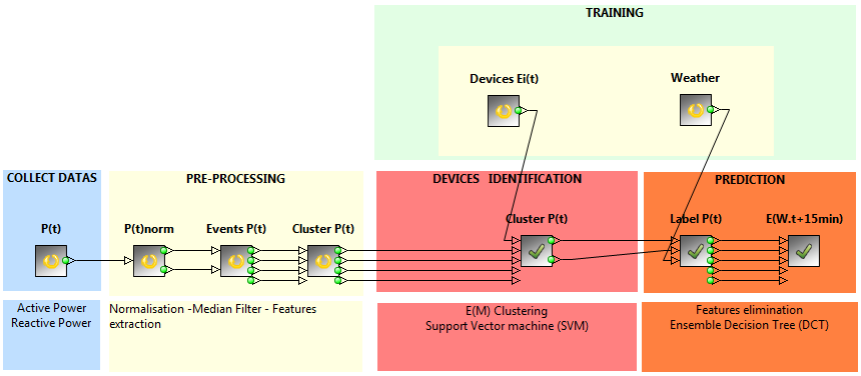

Fig. 1. Steps of the methodology implemented in KNIME

if appliance i is on at t, 0, if appliance i is off at time t for $i = 1...n$. The switch process modulates the power consumption of the individual appliances. A multiphase load with p phases can be modeled as a p-vector in which each component is the load on one phase. The total load p-vector is the sum of the individual appliance load p-vectors for those appliances switched on at any given point in time. This will be a vector function of time which steps in characteristic increments whenever an appliance switches on or off. For i = 1 . . n, let P be the p-vector of the power that the ith appliance consumes when it is operating. The real and imaginary parts for the complex power in the jth component of the vector correspond to the real and reactive power consumed on the jth leg (it is the imaginary part). Then we model $P(t) = \sum a_i(t)P_i + e(t)$. where P(t) is the p-vector as seen at the utility at time t, and e(t) is a small noise or error term. The power consumption varies 20% for reasons external and to the load does not provide an ideal signature [6]. The load admittance, Y(t), can be calculated from the measured power P(t) and the voltage voltage : $P_{Norm}(t) = 220^2 Y(t) = (220/V(t))^2 * P(t)$.
A median Filter is compute to remove high frequencies. The normalized and filtered power by phases are the input vectors to an edge detection algorithm which finds the times of all steplike changes. Many well-known signal processing techniques, such as filtering, differentiating, and peak detection, could be used to find the times at which a signal changes rapidly [10], [9] [11]. The method of visual image processing [18] and an information-based method of [7] could also be adapted to this problem. A key requirement here is that the procedure must not be affected by start-up transients which often accompany steps. Our transient-passing step-change detector first segments the normalized power values into periods in which the power is steady and periods in which it is changing, as indicated with a one-dimensional power signature. A steady period is defined to be one of a certain minimum length in which the input does not vary by more than a specified tolerance (15 Watt or Volt Ampere Reactive(VAR)) in any component. A sequence of time-stamped step change p-vectors is the output combined with active and reactive power variations every 15 minutes.

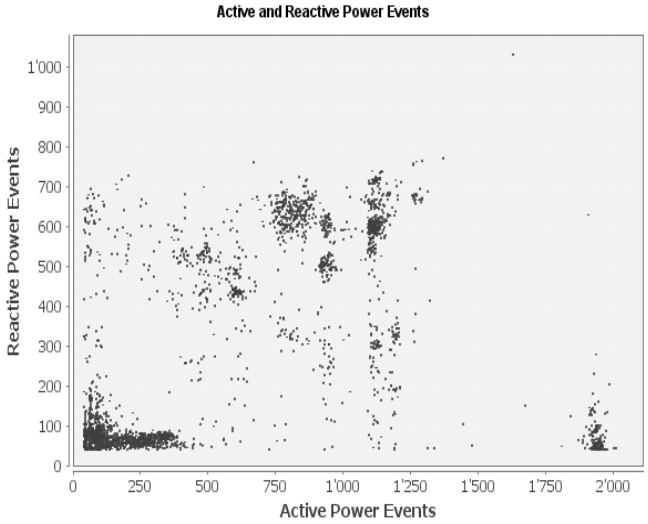

Fig. 2. Active and Reactive events for Phase 1, four months training

B. Training Appliances Models

The ON/OFF model is a good model for most household appliances, such as a toaster, light bulb, or water pump. However, it makes no provision for electrically distinct types of ON states as found in a typical toaster oven (bake/broil/toast), three-way lamp (low/medium/high), or washing machine fill, agitate,spin). That's why we consider the ON/OFF model like associated at a state of a device, clusters in our case. A complex device is divided into components. For example, if the motor and heating element of the dishwasher operate independently, they are learned as two devices, and the energy is appropriately apportioned between the two.
The devices are grouped by family : Heating, Hot Water, Household electrical appliances, Fridge/Freezer,Others. The state of each devices by family are computed by the E(M) algorithm which assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters [17]. E(M) decide how many clusters to create by cross validation. This algorithm terminates when a local optima in the log likelihood function has been found or a maximum number of iterations has been

reached. The function define the acceptance of windows for training upon termination of the EM algorithm. Each one of the generated clusters is a vector of mean values of the selected features forming the centroid of the cluster, and a vector of deviation values associated to the clusters. These vectors can be represented mathematically as: $\mu_{Ci} = [\mu_{f1}, ..., \mu_{fn}]$, and $\sigma_{Ci} = [\sigma_{f1}, ..., \sigma_{fn}]$; where μ_{Ci} is the mean and σ_{Ci} the standard deviation of the cluster i, . Each clusters is referenced to a appropriate label indicating the state of a device. A non-intrusive model must have apriori informations to generalize the devices identification. The heating and the hot water are dependent on the surface of the building and on the number of people.

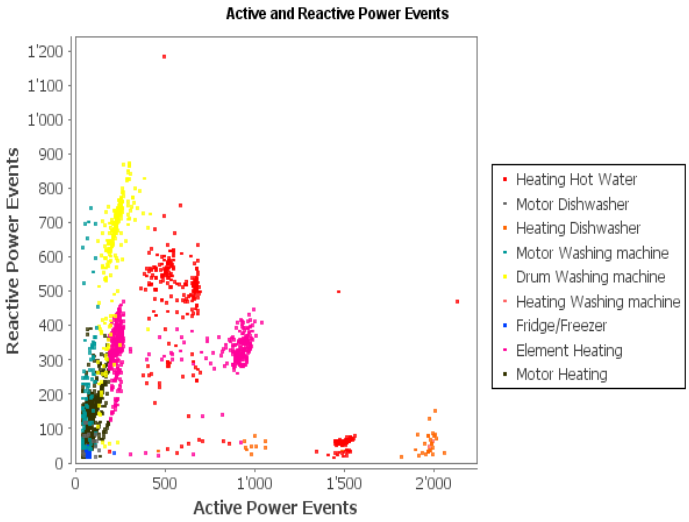

Fig. 3. Active and Reactive events by states of devices

C. Devices Identification

The training set are all the p-vectors of each collected devices. The testing set is the total load curve. The training set are the active and reactive vectors by clusters which represent the different states of devices by family (Heating, Hot Water, Household electrical appliances, Fridge/Freezer,Others). These vectors are the p-active/reactive vectors detected every 15 minutes.They are normalized by the z-score normalization by the formula : $z = (x - \mu/\sigma)$.
In todays machine learning applications, support vector machines (SVM) [21] are considered a must tryit offers one of the most robust and accurate methods among all well-known algorithms. In the active-reactive power space,the problem is not linear. We extend the SVM formulation so that it works in situations where the training data are not linearly separable. A kernel function is used to define a variety of nonlinear relationship between its inputs. For example, besides linear kernel functions, you can define quadratic or exponential kernel functions. Much study in recent years have gone into the study of different kernels for SVM classification and for many other statistical tests. The probability scored threshold enable to assign or not a device with a p vector from the global load curve.
The generalization problem between a different houses y is difficult in particular for the small power devices [15], [16]. In this work, the identification of the Heating, the Hot Water and the household electrical appliances is provide. A class "'Other"' is created for the results below the probability scored threshold.

D. Energy Prediction

The output of devices identification step is a p-vector identified correctly by phases. The sum of energy consumed every 15 minutes and the energy consumed by devices are computed with the time reference and the symbol vector. The day are added and is cut in four parts : [12pm-6am]]6am-12am]]12am-6pm]]6pm-12pm].The temperature is too collected every minutes. The mean by 15 minutes step is computed and added in the data set.
In this work, the output is to check if we have a flexible device started in the next 15 minutes. So we separate our data set in two data set. The first data set is the heating vectors with the temperature. The second data set are the others vectors (Hot Water, Household electrical appliances, Others).
The features elimination step enable to evaluate importance of each input data.In our software,the Backward Feature Elimination loop compute by iteration on data set to find the error rate of each features. The first iteration is executed with all input columns. In the next n - 1 iterations each of the input columns (target column excepted) is left once. Then the node will discard the column that influenced the result the least. Then n - 2 iterations follow where each of the remaining columns is left out once and so on.

A Probabilistic Neural network(NN) [19] [20] and a Ensemble Tree provides for the prediction. The output model describes an ensemble of decision tree models and is applied in the corresponding predictor using a simply majority vote.

E. Accuracy Evaluation

In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function. The sensitivity measures the proportion of actual positives which are correctly identified and is complementary to the false negative rate.

Specificity measures the proportion of negatives which are correctly identified and is complementary to the false positive rate. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

III. Experimental setting

In this section, we describe how our information system collects data, how the parameters are used in our analysis software and how the data set for prediction model.

A. Information System

The schneider electric system provide the active and reactive power by phase from the global electric meter [23]. This smart meter is the PowerLogic Series 800 PM810 of the Schneider Electrical Company [24]. As outputs, we have the amperage, voltage, active and reactive power and energy consumed with in a one second interval per phase. We use the same device to collect the heating. Plugs or Geroco smart meters collect the power of each devices. The protocol of communication used is a Zigbee [26]. It is possible to define activation and dis-activation of theses devices. An integrated pre-programmable code enables the recognition of the different devices.
The data travels through a modbus communication.These data are stored on mini-pc in csv files and to send on server in HES SO in Sion. We use the open source data analysis software KNIME [25] to connect the database and process the data.

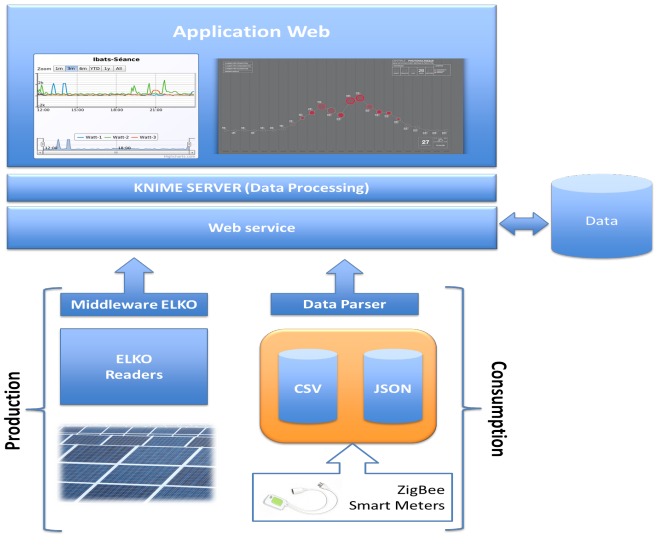

Fig. 4. I-BAT information system architecture based No-SQL databases

B. Data set

The aim is to maximize the training set for the devices identification to not collect directly in the test house. In this paper, the training set represent one year data by devices collected from the test houses or our research institute. The prediction objective is to detect a level of flexibility during the next 15 minutes. The winter season represent the good period to analyze the heating and it is the difficult period for the producers and suppliers energy. Because we does not have air conditioning during the summer season in our test houses.
For a prediction, an history needs to be created. Data are available but cannot be associated together. This history is created through a lag process. This will create a new line with value in the past. For example, for a specific node at a precise time, the outside temperature (6 hours before) will be added as additional columns. Because of the difference between the number of Heating active-reactive vectors and the no heating active-reactive vectors, a Bootstrap sampling is done. This will equalize both samples and the accuracy of the model is improved.
Four month during the 2012 Winter season represent the data set. For the prediction, 60% of them are used for the training process and 40% of them for the test.

IV. Classification results

The results are presented for a test house. A winter season represent the testing set for the identification step. One year data set by states of devices represent the training data set. After the identification by the support vector machine, the active-reactive power vectors with the timestamps are the input for the prediction step compute by a ensemble of a decision tree. The Hot Water, the heating,the Household electrical appliances, the drum of a washing machine are identify. A class "'Other"' is created for the results below the probability scored threshold. The support vector machine is the best algorithm to classify the electrical appliances in the active reactive power plan and provide an identification of 95% +/- 2. After the identification, an evaluation of each features is computed.This elimination improves the quality of the prediction by excluding some noisy features. Finally, two models are evaluate (the hot Water and the heating model) which represent the flexible energy in a residential building. To evaluate the importance of each features, the feature elimination is done over the input parameters. Between the probabilistic neural network and an ensemble of decision tree enable the prediction.The results are suitable with the ensemble of decision tree and only these results are presented. For the hot water prediction model, 29 features are necessary after the feature elimination. It is split into three categories (the day period, the total energy consumed and the hot water energy consumed). In the Figure 5, the ROC Curve represent the best results per inputs variables for the hot water prediction model : the energy consumed in the last 15 minutes, the total energy consumed in the last 15 and 30 minutes. The algorithm needs 225 minutes of history for the total energy consumed and 195 minutes of history for the hot water energy consumed. All the results are presented in the table I.

Variables	Area	Variables	Area
Hot Water energy t-1	0.77	Total Energy W.15min t -1	0.928
Hot Water energy t-2	0.741	Total Energy t -2	0.884
Hot Water energy t-3	0.685	Total Energy t -3	0.84
Hot Water energy t-4	0.645	Total Energy t -4	0.802
Hot Water energy t-5	0.689	Total Energy t -5	0.786
Hot Water energy t-6	0.652	Total Energy t -6	0.749
Hot Water energy t-7	0.563	Total Energy t -7	0.687
Hot Water energy t-8	0.587	Total Energy t -8	0.68
Hot Water energy t-9	0.634	Total Energy t -9	0.611
Hot Water energy t-10	0.55	Total Energy t -10	0.556
Hot Water energy t-11	0.461	Total Energy t -11	0.571
Hot Water energy t-12	0.576	Total Energy t -12	0.556
Hot Water energy t-13	0.518	Total Energy t -13	0.554
Day Period	0.574	Total Energy t -14	0.543
		Total Energy t -15	0.515

TABLE I. AREA UNDER CURVE FOR THE HOT WATER PREDICTION MODEL

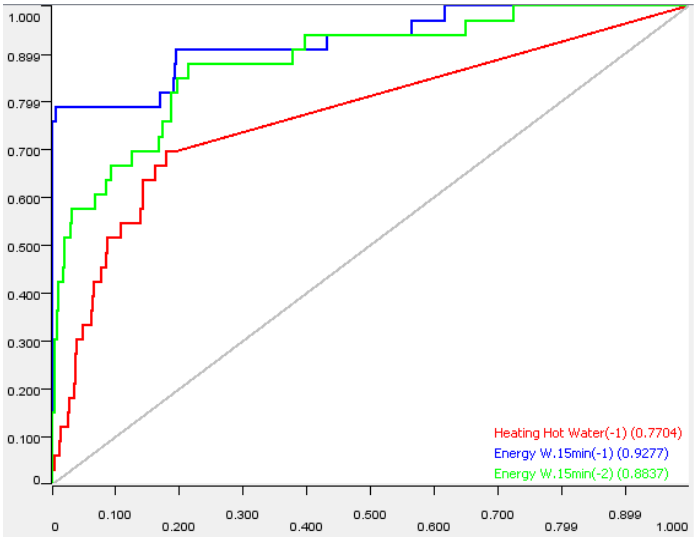

Fig. 5. Courbe ROC for the hot water prediction model with the best input variables

For the heating model, 65 features are necessary after the feature elimination. It is split into four categories (the day period, the hour, the total energy consumed and the heating energy consumed and the hours).
In the Figure 6, the ROC Curve represent the best results per inputs variables for the heating model : the outside temperature in the last 225 and 240 minutes, the total energy consumed in the last 45 and 60 minutes. In fact, the impact of the outside temperature on the heating depends of the inertial mass of the building.
For our test house, the algorithm needs 300 minutes of history for the outside temperature and 195 minutes of history for the total energy consumed. All the results for this model are presented in the table II.

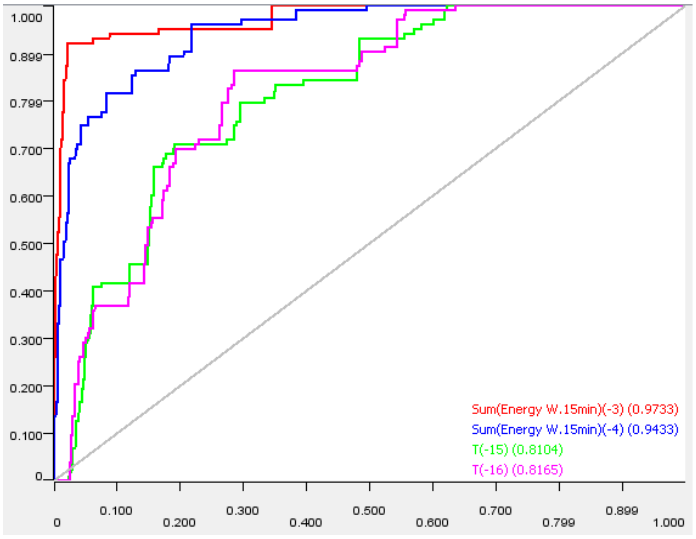

Fig. 6. Courbe ROC for the heating prediction model with the best input variables

Variables	Area	Variables	Area
Outside Temperature t-1	0.714	Total Energy W.15min t -1	0.905
Outside Temperature t-2	0.723	Total Energy t -2	0.958
Outside Temperature t-3	0.745	Total Energy t -3	0.973
Outside Temperature t-4	0.742	Total Energy t -4	0.943
Outside Temperature t-5	0.736	Total Energy t -5	0.904
Outside Temperature t-6	0.75	Total Energy t -6	0.834
Outside Temperature t-7	0.767	Total Energy t -7	0.77
Outside Temperature t-8	0.771	Total Energy t -8	0.732
Outside Temperature t-9	0.77	Total Energy t -9	0.718
Outside Temperature t-10	0.772	Total Energy t -10	0.709
Outside Temperature t-11	0.791	Total Energy t -11	0.705
Outside Temperature t-12	0.795	Total Energy t -12	0.656
Outside Temperature t-13	0.792	Total Energy t -13	0.621
Outside Temperature t-14	0.799	Day Period	0.694
Outside Temperature t-15	0.81	Hour	0.756
Outside Temperature t-16	0.816		
Outside Temperature t-17	0.808		
Outside Temperature t-18	0.802		
Outside Temperature t-19	0.804		
Outside Temperature t-20	0.805		

TABLE II. AREA UNDER CURVE FOR THE HEATING PREDICTION MODEL

V. Conclusions

This work presented a non intrusive methodology to identify the devices and to predict the flexible energy for the next 15 minutes. The devices identified are : Heating, Hot Water, Household electrical appliances and Others. A model on/off enable to separate the states of these devices. The active and reactive power are the variable to identify the devices from the global load curve. However, The small power devices, generally below 250 Watt, are not identify in this methodology because these devices are very different between the houses or not present (Computer, Light...) . The support vector machine provide the identification around 94% in the active-reactive power plan. The prediction provide the flexible energy for the next 15 minutes : the detection of the heating and the hot water.
For the heating prediction model, the outside temperature in the last 225 and 240 minutes and the total energy consumed in the last 45 and 60 minutes are the variables the most important for our test house. For the hot water prediction model,the energy consumed in the last 15 minutes, the total energy consumed in the last 15 and 30 minutes are the variables the most important for our test house. In our future works, the parameters buildings like the surface area of the windows, the humidity, the wind will be added. It is possible for the next step to control this devices in order to balance the production by the photovoltaic panel with the consumption.

VI. acknowledgements

This research is a part of the I-BAT project and was financed by The Ark Energy. It is too an element of a thesis between the Mines Telecom Albi with the RAPSODEE research institute and the HES SO Valais with the information system research institute in Sierre.

References

[1] Thomas F. Garrity, VP, Sales and Business Development, Siemens Power available. URL = https://w3.energy.siemens.com/cms/us/whatsnew/Documents/Getting20%Smart$_{G}arrity.pdf$

[2] IEEE (2012) OFEN, Statistique suisse de l'lectricit en 2012. *The Edison Foundation Institute for Electric Efficiency*. URL = http://www.strom.ch/fr/dossiers/graphiques-electricite.html.

[3] Reported by the Gerson Lehrman Group. URL = http://e360.yale.edu/content/digest.msp?id=2252.

[4] Zoha, A., Gluhak, A., Imran, M.A., Rajasegarar, S. (2012) Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey. *Sensors*. vol. 12, no. 12, pp. 16838-16866.

[5] IEEE (2006) Utility-scale smart meter deployments, plans and proposals. *The Edison Foundation Institute for Electric Efficiency*.

[6] Hart, G.W. (1992) Nonintrusive appliance load monitoring. *Proceedings of the IEEE*. vol. 80, no. 12, pp. 1870-1891.

[7] Farinaccio, L. (1999) The disaggregation of whole-house electric load into the major end-uses using a rule-based pattern recognition algorithm. *Concordia University*, PhD Thesis.

[8] Marceau, M.L., Zmeureanu, R. (2000) Nonintrusive load disaggregation computer program to estimate the energy consumption of major end uses in residential buildings. *Energy Conversion and Management*. vol. 41, no. 13, pp. 1389-1403.

[9] Srinivasan, D. and Ng, W. S. and Liew, A.C. (2006) : Neural-network-based signature recognition for harmonic source identification, Vol. 21,no. 1,pp. 398-405.

[10] Patel, S., Robertson, T., Kientz, J., Reynolds, M., Abowd, G. (2007) At the Flick of a Switch: Detecting and Classifying Unique Electrical Events on the Residential Power Line. *UbiComp 2007: Ubiquitous Computing. Lecture Notes in Computer Science*. Vol. 4717, Springer Berlin Heidelberg, pp. 271-288.

[11] El Guedri, Mabrouka (2009) Caractérisation aveugle de la courbe de charge électrique: Détection, classification et estimation des usages dans les secteurs résidentiel et tertiaire

[12] Luan, S. W., Teng, J. H., Chan, S. Y., Hwang, L. C. (2009). Development of a smart power meter for AMI based on ZigBee communication. *In Power Electronics and Drive Systems, PEDS 2009. International Conference on*. pp. 661-665, IEEE.

[13] Froehlich, J., Larson, E., Gupta, S., Cohn, G., Reynolds, M., Patel, S. (2011). Disaggregated end-use energy sensing for the smart grid. *Pervasive Computing, IEEE*. Vol. 10, no. 1, pp. 28-39.

[14] T. Cover,P. Hart : Nearest Neighbor Pattern Classification(1967), journal=The Edison Foundation Institute for Electric Efficiency,*The Edison Foundation Institute for Electric Efficiency* pp. 21-27.

[15] L. Dufour,D. Genoud, B. Ladevie,JJ. Bezian,(2014) Verification method for home electrical signal disaggregation, *ESIOT, IEEE*

[16] L. Dufour,D. Genoud,B. Ladevie, JJ. Bezian,(2014) Identification method for home electrical signal disaggregation, *SET , IEEE*

[17] Yihua Chen and Maya R. Gupta, EM Demystified: An Expectation-Maximization Tutorial,UWEE Technical Report,2010

[18] A. Blake and A. Zisserman,Visual Reconstruction, *MA : MIT Press*,1987

[19] Simon . H ,Neural networks: a comprehensive foundation,Prentice Hall,1999

[20] Lane, Stephen H and Flax, Marshall G and Handelman, David A and Gelfand, Jack J,(1990)Multi-layer perceptrons with B-spline receptive field functions,pp 684-692 *Proceedings of the 1990 conference on Advances in neural information processing systems 3*,

[21] Fast Training of Support Vector Machines using Sequential Minimal Optimization, URL = http://research.microsoft.com/en-us/um/people/jplatt/smo-book.pdf.

[22] John Shafer, Rakesh Agrawal, Manish Mehta,SPRINT, A scalable parallel classifier for data URL = http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.152rep.pdf.

[23] Power meter series 800 data sheet,URL = http://www2.schneider-electric.com/resources/sites/ SCHNEIDER ELECTRIC/content/live/FAQS/140000/FA140345/enUS/PM80020DataSheet.pdf.

[24] Power meter series 800 user guide,URL = http://download.schneider-electric.com/files?p File Id= 27600253p File Name=63230-500-225A2 PM800 User Guide EN.pdf.

[25] Knime,URL = http://knime.com/.

[26] Ecowizz,URL = https://www.ecowizz.net/.